Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 3443–3446, June 1984
Genetics

Strategies for multilocus linkage analysis in humans

(DNA polymorphism/recombination/linkage detection/location score/gene mapping)

G. M. LATHROP*, J. M. LALOUEL*, C. JULIER*†, AND J. OTT‡

*Laboratoire d'Anthropologie biologique, Université de Paris 7, 2 place Jussieu, 75005 Paris, France; †Institut de Pathologie Moléculaire, Institut National de la Santé et de la Recherche Médicale, U 129, CHU Cochin, 75014 Paris, France; and ‡City Statistics Office, Napfgasse 6, 8001 Zürich, Switzerland

Communicated by Jean Dausset, February 6, 1984

ABSTRACT **The increasing number of DNA polymorphisms characterized in humans will soon allow the construction of fine genetic maps of human chromosomes. This advance calls for a reexamination of current methodologies for linkage analysis by the family method. We have investigated the relative efficiency of two-point and three-point linkage tests for the detection of linkage and the estimation of recombination in a variety of situations. This led us to develop the computer program LINKAGE to perform multilocus linkage analysis. The investigation also enables us to propose a method of location scores for the efficient detection of linkage between a disease locus, or a new genetic marker, and a linkage group previously established from a reference panel of families. The method is illustrated by an application to simulated pedigree data in a situation akin to Duchenne muscular dystrophy. These results show that considerable economy and efficiency can be brought to the mapping endeavor by resorting to appropriate strategies of detecting linkage and by constructing the human genetic map on a common reference panel of families.**

Various methods are available for investigating relationships between genetic loci in humans (1). On one hand, methods such as somatic cell hybrids provide physical assignments of loci on human chromosomes, contributing directly to the physical map. On the other hand, linkage analysis by the family method leads to the construction of a genetic map, where the distances between loci depend on both the extent of physical separation and the rate of occurrence of crossing-over. The latter approach allows the detection of linkage and the estimation of recombination rates and, thus, yields estimates of genetic parameters relevant for investigations of modes of inheritance, resolution of etiological heterogeneity, calculation of genetic risks, and genetic mapping.

Just as molecular hybridization has given a new power to methods for physical assignments (1), the new wealth of DNA polymorphisms (2, 3) will elicit the development of new strategies for linkage analysis by family methods. When only about 30 genetic markers were available at arbitrary locations, affording a very partial coverage of the human genome, a natural approach for the detection of linkage between a disease locus and a battery of markers consisted in the pairwise analysis of the disease phenotype and each marker in turn. Two-locus linkage analysis by the now classical method of lod-scores (4) or related techniques was originally restricted to simple Mendelian traits and nuclear families; later it was extended to complex phenotypes and general pedigrees through the development of appropriate algorithms and computer programs (5–7).

More than 200 DNA polymorphisms have been defined in recent years (8), and there is no doubt that the number required to span the human genome (2, 9) will be reached soon. This inevitably raises questions regarding the relative merits of two-point and multipoint linkage analysis. Although the advantages of multipoint tests, as opposed to pairwise tests, seems generally intuitive (10), a systematic investigation is necessary before new approaches can be proposed.

Need for a multilocus analysis is evident for the calculation of genetic risks when several linked markers are available; otherwise there would be no general way of combining pedigree calculations involving each marker singly. For detection of linkage, estimation of recombination, and construction of genetic maps, the merit of multipoint tests has yet to be established. Although Meyers *et al.* (11) considered three-point tests in restricted situations, most procedures for estimation of recombination and genetic mapping in humans have been based on the assumption that results from independent two-point linkage tests are combined (12–14).

The determination of a genetic map from results of linkage analyses requires assumptions about the mathematical relationships between map distance, expressed in units of crossing-over, or morgans, and recombination frequency, thus defining a mapping function. This relation is complex because recombination results from an odd number of points of exchange between loci, and evidence points to their nonindependence—i.e., interference in crossing-over (15). Various mapping functions have been proposed embodying specific assumptions regarding interference (15). Statistical methods have been proposed that assume a mapping function or a specific process of chiasma formation (12–14) or that infer a genetic map solely from the rank-order constraints implied by pairwise recombination estimates (16).

As the genetic map is developed, it is likely to have an impact on strategies used to assign a new genetic marker or disease locus to a preexisting map; increased efficiency can be achieved by the use of previously estimated recombination rates between markers in multipoint tests. An advantage of this approach can be seen by noting that the probability a family is informative for linkage at one or more locus increases with the number of loci studied, and that each informative locus may provide information on the location of the new marker or disease relative to the existing map.

In this paper, we examine the relative efficiency of two-point and three-point linkage tests under various conditions. We then consider the relevance of multilocus analysis for the construction of genetic maps and propose a strategy for the detection of linkage and the addition of a new locus to a known genetic map. The computer program package LINKAGE has been developed for this investigation and is available from the authors on request.

METHODS

Three-Point Linkage. Suppose that data are available for three linked loci A, B, and C in the same families, and denote the three recombination rates as θ_{AB}, θ_{AC}, and θ_{BC}. The classical approach consists in analyzing each pairwise combination of loci by computing a lod-score (i.e., the decimal logarithm of the odds for various values of recombination against

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

no linkage) and taking as the estimate of recombination that value for which the lod-score is maximum. The gene order may be inferred by inspection of the estimated recombination rates. However, this approach provides no means of assessing evidence for alternate orders and may not use the data fully efficiently.

An alternative consists in analyzing the three loci jointly; several strategies then can be entertained. No assumptions on gene order and interference are needed if one defines and estimates the probabilities of the following recombination events (15): (*i*) recombination between both *AB* and *BC*; (*ii*) recombination between *AB* and no recombination between *BC*; (*iii*) no recombination between *AB* and recombination between *BC*. Denoting these three probabilities p_1, p_2, and p_3, respectively, we have the simple relationships: $\theta_{AB} = p_1 + p_2$, $\theta_{BC} = p_1 + p_3$ and $\theta_{AC} = p_2 + p_3$. The estimates yield directly the most likely gene order.

If a specific mapping function is assumed, only two recombination rates need be estimated; the recombination between flanking markers can be expressed in terms of the rates within each contiguous segment and an interference parameter specified by the mapping function. However, in situations prevailing in humans, little information may be available to detect interference. For instance, with interference at the Kosambi level (17) (i.e., intermediate between complete and no interference) and recombination of 15–20% between adjacent loci, over 850 offspring observations are required in a phase-known triple backcross to obtain a power of 0.80 in order to detect interference; more than 4100 offspring in families with two children are needed when the phase is unknown. For smaller recombination rates, even larger samples are required (unpublished data). It follows that the choice of a mapping function may not be critical for data from humans; therefore, one may assume any meaningful function. For instance, we may suppose that there is no interference—i.e., that recombination events are independent in different segments of the chromosome. Linkage analysis under this assumption requires consideration of all possible gene orders, but within each, only the recombination rates in the two contiguous segments need be estimated; thus, if the given order is *ABC*, we have:

$$\theta_{AC} = 1 - (1 - \theta_{AB})(1 - \theta_{BC})$$

Another situation arises if we investigate linkage between a certain locus and a pair of loci when recombination between the latter has been estimated from other evidence to a reasonable degree of accuracy. This will be the case when a disease locus or any other rare trait is analyzed with respect to a preestablished linkage group. Under the assumption of no interference or with the use of a specified mapping function, the problem reduces to the estimation of a single recombination parameter.

Relative Efficiency of Three-Point and Two-Point Linkage Tests. The relative merits of two-point and three-point linkage analyses can be investigated in terms of the statistical efficiency of the estimates of recombination. Efficiency is defined as the ratio of the variances of the two-point and three-point recombination estimates for given loci and study design. The estimate with the smaller variance is deemed more efficient as it provides more precise knowledge of relative gene locations.

The relative efficiency depends on the recombination rates, the mode of inheritance of the traits, and the type of family data used. For simplicity, we will consider the minimal situation where families consist of two parents and two offspring in which phenotype information is available for the three loci on all individuals. For a sufficiently large sample of randomly chosen families, the variances of the maximum likelihood estimates of recombination can be obtained by using standard likelihood theory (18). Briefly, the expected values of the second derivatives of the likelihood function are calculated, and the resulting information matrix is inverted to yield the variance-covariance matrix of the estimates.

Fig. 1 documents results for various modes of inheritance at a central locus, *B*, and two codominant flanking loci, *A* and *C*, under two conditions: $\theta_{AB} = \theta_{BC} = 0.1$ and $\theta_{AB} = 0.1$, $\theta_{BC} = 0.05$. The efficiency of two-point linkage analysis is given relative to three possible strategies of multilocus analysis: (*i*) when no assumptions are made regarding recombination and interference; (*ii*) when the recombination between flanking loci is known; and (*iii*) the previous situation under the additional assumption of no interference. The relative efficiencies are similar for the dominant, codominant, and recessive situations considered but vary with the recombination rates and the strategy selected. The gain in information is most substantial when the recombination rate between the flanking markers has been previously established and no interference is assumed: the efficiency of the three-point estimate relative to the two-point estimate may be increased as much as 4-fold. We verified that further increase in efficiency can be achieved by selecting families on the basis of parental mating types (unpublished data). Selection of families to be informative at two test loci may be a particularly effective strategy as shown in the example below.

Adding a Locus to a Genetic Map. Our results indicate that multilocus analysis can substantially improve the efficiency of linkage analysis and, hence, the accuracy of the inferred genetic map. For genetic diseases and other rare traits, these considerations have important implications for mapping strategies. An effective approach would consist in first establishing the genetic map of the test markers to sufficient accuracy on a panel of control families. Since the data for the disease locus will generally be much more limited than for the test markers, we may assume that the genetic map of the

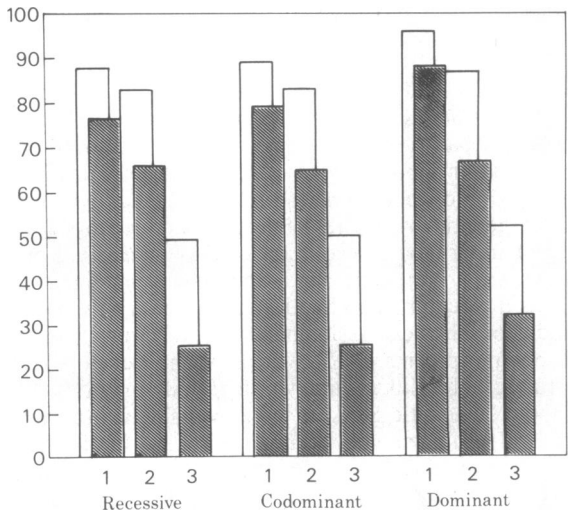

FIG. 1. Relative efficiency of two-point vs. three-point linkage tests for families with both parents and two children studied. Three modes of inheritance are considered for the central locus (*B*): recessive, codominant, and dominant. For recessive and dominant traits, the population prevalence of affected individuals is assumed to be 0.0001, and families are selected to include at least one affected offspring. For the codominant case, gene frequencies are taken to be 0.5. Two diallelic flanking loci (*A* and *C*) are assumed, with gene frequencies of 0.5. Recombination values are: $\theta_{AB} = \theta_{BC} = 0.1$ (white histogram); $\theta_{AB} = 0.1$, $\theta_{BC} = 0.05$ (shaded histogram). Efficiencies are reported for θ_{AB} in both instances for the three strategies: joint estimation of three recombination rates (histograms 1); θ_{AC} known (histograms 2); θ_{AC} known and assuming no interference (histograms 3).

latter is known exactly. The joint segregation at the disease and at test loci can then be expressed in terms of a single parameter: the location of the disease locus relative to the genetic map of the selected test markers. The same strategy would seem appropriate for the detection of linkage between a new marker locus and a linkage group previously established on the panel of control families.

This approach has several advantages: multilocus analysis is the only way of accounting for the nonindependence of the recombination estimates; it affords increased precision of the estimated location of a new locus on the genetic map; and, under the assumption of no interference, it also reduces the problem of estimating multiple parameters to that of estimating a single parameter. In the following example, a convenient graphical method of location scores with some analogy to the classical lod-score curve is proposed. This method allows an easy combination of results from different studies.

EXAMPLE

The following example will illustrate the method of location scores. Data were simulated for a situation akin to Duchenne muscular dystrophy. We consider two diallelic loci, *A* and *C*, on either side of the Duchenne Locus, *D*, with gene frequencies of the rarer allele of 0.13 and 0.32, respectively, each recombining with Duchenne at the rate of 0.17 (19, 20). For greater generality, we have assumed a third diallelic locus, *B*, located between loci *D* and *C*, with a frequency of 0.20 for the rarer allele and a recombination rate of 0.10 with the Duchenne locus. Thus, the order of the loci is *A-D-B-C*. Parameters at the Duchenne locus were assigned in accordance with reported figures (21): mutation rate of 0.0001, gene frequency of 0.0003, and logarithm of creatine kinase of 1.57 ± 0.24 and 2.10 ± 0.41 in normal and carrier women. The fixed family structure of Fig. 2 has been assumed; because individual 4 has an affected son and an affected brother, she is an obligate carrier. Random assignments of alleles and of creatine kinase values for the females of the bottom generation were made in accordance with these figures and Mendelian segregation, assuming linkage equilibrium. Two series of 10 pedigrees were selected according to one of two criteria: the obligate carrier is heterozygote for at least one or at least two of the three test markers.

Without interference, recombination between contiguous segments such as *A-B* and *B-C* satisfies the equation $\theta_{AC} = \theta_{AB}\cdot(1 - \theta_{BC}) + (1 - \theta_{AB})\cdot\theta_{BC}$. The relationships between genetic distance, *d*, and recombination, θ, are given by Haldane's mapping function $d = -[\ln(1 - 2\theta)]/2$ and its inverse $\theta = [1 - \exp(-2d)]/2$ (22). Thus, the recombination rates we have assumed, $\theta_{AD} = \theta_{DC} = 0.17$ and $\theta_{DB} = 0.10$, lead to $\theta_{AB} = 0.24$, $\theta_{AC} = 0.31$, and $\theta_{BC} = 0.09$. The genetic distances between loci can be expressed in terms of map location, *w*. With locus A chosen as the origin (i.e., $w_A = 0.0$) the application of Haldane's function to the assumed recombination rates leads to the following map locations: $w_D = 0.21$, $w_B = 0.32$, and $w_C = 0.42$. Since genetic distances are additive over contiguous segments, $d_{AC} = d_{AD} + d_{DB} + d_{BC} = 0.21 + 0.11 + 0.10 = 0.42$.

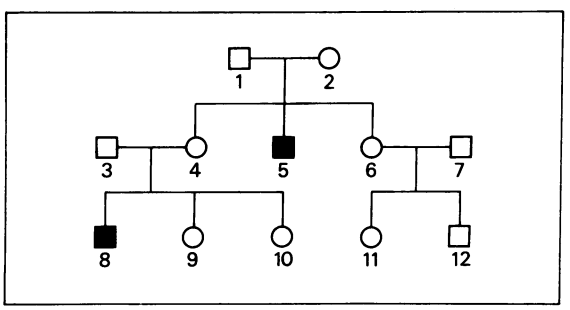

FIG. 2. Pedigree structure used for the simulation of Duchenne muscular dystrophy and three marker loci. Creatine kinase measurements were simulated for females in the last generation. Affected males are denoted by shaded squares.

For each set of data, the genetic locations of loci *A*, *B*, and *C* have been kept fixed to their true values, while varying the genetic location of locus *D* around its true value over a range large enough to encompass free recombination on either side of the cluster *A*, *B*, and *C*. For each new value of w_D, the likelihood for w_D was found with the program LINKAGE by computing the probability of the joint segregation at the four loci in these pedigrees as a function of w_D. In Fig. 3 we report twice the natural logarithm of the odds, the location score $l(w_D)$, for any w_D against a value large enough to imply no linkage. Note that while $l(w_D)$ is a function of a parameter of genetic location rather than recombination, a lod-score scale can be obtained by dividing $l(w_D)$ by $2\ln(10) \simeq 4.6$.

Support for linkage of Duchenne with the cluster *A*, *B*, and *C* can be assessed by comparing the maximum value of $l(w_D)$ to a χ^2 variate with one degree of freedom. This test is highly

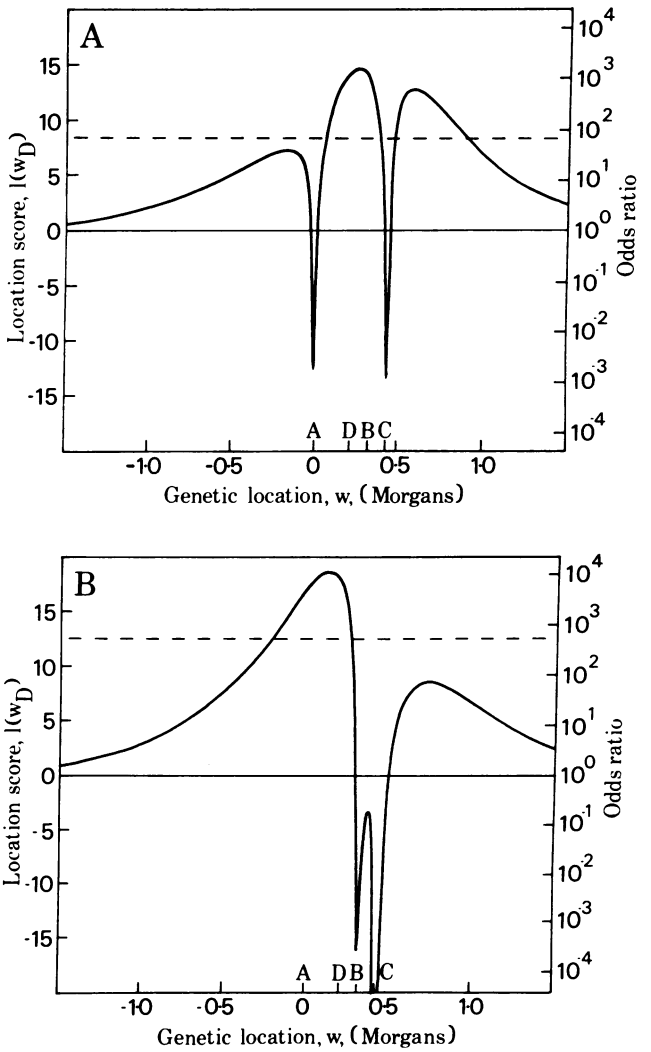

FIG. 3. Likelihood of the map location of a simulated disease locus (*D*) with respect to three linked loci (*A*, *B*, and *C*) for 10 replicates of the pedigree in Fig. 2. Data selected so that individual 4 of Fig. 2 is heterozygote for at least one test marker (Fig. 3*A*) or at least two test markers (Fig. 3*B*). Horizontal axis, genetic distance *w* from locus *A*; right vertical axis, odds ratio for location of locus *D* at w_D vs. *D* at infinite distance (implying no linkage); left vertical axis, location score $l(w_D)$ defined as twice the natural logarithm of the odds ratio; ---, lower limits for locations with odds < 20:1 relative to the overall maximum.

significant in both cases, yielding 14.4 and 18.1 for the selection on one and two informative loci. The corresponding values on a lod-score scale are 3.1 and 3.9. By contrast, for the selection on one locus, the pairwise maximum lod-score are 0.6, 1.8, and 1.0 for the pairs (*A*, *D*), (*B*, *D*), and (*C*, *D*), none of which reaches the value of 3.0 often used, somewhat arbitrarily, as the critical level for significant evidence of linkage. The pairwise maxima occur at different points of the genetic map. Pairwise lod-score curves cannot be summed to produce an overall test because they are nonindependent. For both sets of data, the highest mode of $l(w_D)$ occurs between loci *A* and *B*, with $l(w_D) = 14.4$ at $w_D = 0.24$ in Fig. 3*A* and $l(w_D) = 18.1$ at $w_D = 0.15$ in Fig. 3*B*. This corresponds to maximum likelihood estimates for θ_{AD} of 0.19 and 0.13, respectively, close to the true value of 0.17.

In both cases, $l(w_D)$ is clearly multimodal, as values of w_D become extremely unlikely whenever they bring the Duchenne locus close to a test locus for which affected males give evidence of recombination; this is so for loci *A* and *C* in Fig. 3*A* and for loci *B* and *C* in Fig. 3*B*. Antimodes are not observed for locus *B* in Fig. 3*A* and for locus *A* in Fig. 3*B*; this is because there is no apparent recombination between these loci in affected males. Unless the disease locus is completely linked to a test locus, antimodes will occur at the location of any test locus in large data sets, with true discontinuities in the likelihood function whenever the loci considered are fully penetrant without mutation. With such discontinuities or near discontinuities, a solution to the mapping problem by maximum likelihood methods raises complex statistical and numerical issues.

Because of the shape of the likelihood function, it would be misleading to associate confidence intervals to the standard errors of such estimates since these only reflect the curvature of $l(w_D)$ in the immediate neighborhood of a local maximum. This difficulty is not particular to the approach discussed here; it appears to be inherent to the mapping problem when using parametric statistical methods. Nevertheless, one can tentatively assess support in favor of various gene orders by comparing the several maxima of $l(w_D)$. In Fig. 3*A*, the relative odds of orders *A-D-B-C*, *A-B-C-D*, and *D-A-B-C* are 35:13:1. Fig. 3*B* illustrates the increased information on order provided by selecting families informative for at least two marker loci. In this case, the relative odds of orders *A-D-B-C*, *A-B-C-D*, and *A-B-D-C* are 51,534:403:1.

DISCUSSION

An efficient strategy is now emerging for the characterization of new DNA polymorphisms. This starts with the investigation of chromosome-specific libraries and the determination of the physical location of potentially useful probes by hybridization experiments against a panel of cell lines with chromosome rearrangements or related techniques. When the relative location of a probe with regards to previously mapped genetic markers appears likely to provide evidence of linkage, polymorphisms may be sought by restriction digests of DNA from parents or other founding members of a reference panel of families. When a new polymorphism is revealed that seems of value for genetic analysis, its pattern of inheritance can be confirmed by study of the offspring in informative families. Linkage can rapidly and efficiently be detected by the method of location scores. Recombination rates can be reestimated from this panel through multipoint tests in regions where significant new observations have been obtained.

The advantage of assembling an appropriate reference panel of families on which the detection of polymorphisms and the formal genetic analysis would be carried out seems overwhelming. The linkage relationships between a marker and a genetic map can be established by analysis of this panel. When family data are later submitted to linkage analysis, the previously established genetic map may be put to effective use by the joint analysis of the disease phenotype and a battery of linked markers. This is most appropriate when other evidence has led to the assignment of a locus to a given chromosome or chromosome segment. But it also will bring some order to the "shotgun" approach to linkage analysis by allowing the selection of markers to characterize a linkage group and, thereafter, scanning each linkage group in turn by the method of location scores. Combined with an orderly selection of genetic markers within and among linkage groups at different stages of the investigation, it will considerably reduce the number of tests required to detect linkage.

The endeavor of mapping the human genome naturally lends itself to a cooperative effort among investigators. A preliminary step in this direction has been taken with creation of the Human Polymorphism Study Center (CEPH, J. Dausset director, Hopital St Louis, Paris). The goal of this association is to make freely available to the scientific community samples of DNA extracted from controlled, perennial cell lines from a reference panel of families. This association also will maintain and update a data base of the results obtained on this panel, thereby providing ready access for all researchers to current knowledge on genetic markers and the status of the genetic map.

We thank S. Wood and R. H. Ward for helpful discussions. We are grateful to J. C. Kaplan, C. Junien, and D. Cohen for motivation and encouragement. Computing facilities were provided by the Institut National d'Etudes Démographiques, Paris, France.

1. D'Eustachio, R. & Ruddle, F. H. (1983) *Science* **220,** 919–924.
2. Botstein, D., White, R., Skolnick, M. & Davis, R. (1980) *Am. J. Hum. Genet.* **32,** 314–331.
3. Housman, D. & Gusella, J. (1981) in *Genetic Research Strategies for Psychobiology and Psychiatry*, eds. Gershon, E., Matthysse, S., Breakefield, X. O. & Ciaranello, R. D. (Boxwood, Pacific Grove, CA), pp. 17–24.
4. Morton, N. E. (1955) *Am. J. Hum. Genet.* **7,** 277–318.
5. Ott, J. (1974) *Am. J. Hum. Genet.* **26,** 588–597.
6. Hasstedt, S. & Cartwright, P. (1979) *University of Utah Department of Medical Biophysics and Computing Technical Report 13* (Salt Lake City, UT).
7. Morton, N. E. & Lalouel, J. M. (1981) *Hum. Hered.* **31,** 3–7.
8. Human Gene Mapping 7th International Workshop (1983) *Cytogenet. Cell Genet.*, in press.
9. Lange, K. & Boehnke, M. (1982) *Am. J. Hum. Genet.* **34,** 842–845.
10. Edwards, J. H. (1982) *Cytogenet. Cell Genet.* **32,** 43–51.
11. Meyers, D. A., Conneally, P. M., Lovrien, E. W., Magenis, R. F., Merritt, A. D., Norton, J. A., Palmer, C. G., Rivas, M. L., Wang, I. & Yu, P. L. (1976) *Birth Defects: Original Article Series* (The National Foundation, New York), pp. 335–339.
12. Renwick, J. H. & Bolling, D. R. (1971) *J. Med. Genet.* **4,** 399–406.
13. Sturt, E. (1975) *Ann. Hum. Genet.* **39,** 255–260.
14. Rao, D. C., Keats, B. J. B., Lalouel, J. M., Morton, N. E. & Yee, S. (1979) *Am. J. Hum. Genet.* **31,** 680–696.
15. Bailey, N. T. J. (1961) *Introduction to the Mathematical Theory of Genetic Linkage* (Clarendon, Oxford), pp. 137–168.
16. Lalouel, J. M. (1977) *Heredity* **38,** 61–77.
17. Kosambi, D. D. (1944) *Ann. Eugenics* **12,** 172–175.
18. Rao, C. R. (1973) *Linear Statistical Inference and Its Applications* (Wiley, New York), pp. 351–374.
19. Murray, J. M., Davies, K. M., Harper, P. S., Meredith, L., Mueller, C. R. & Williamson, R. (1982) *Nature (London)* **300,** 69–71.
20. Davies, K., Pearson, P., Harper, P., Murray, J., O'Brien, T., Sarfazazi, M. & Williamson, R. (1983) *Nucleic Acids Res.* **11,** 2303–2312.
21. Morton, N. E. & Lalouel, J. M. (1979) *Birth Defects: Original Article Series* (The National Foundation, New York), Vol. 15, pp. 9–24.
22. Haldane, J. B. S. (1919) *J. Genet.* **8,** 299–309.